



The L_1 -consistency of Dirichlet mixtures in multivariate Bayesian density estimation[☆]

Yuefeng Wu^{a,*}, Subhashis Ghosal^{b,*}

^a Department of Biological Statistics and Computational Biology, Cornell University, United States

^b Department of Statistics, North Carolina State University, United States

ARTICLE INFO

Article history:

Received 14 April 2009

Available online 19 June 2010

AMS 2000 subject classifications:

62G07

62G20

Keywords:

Posterior consistency

Dirichlet process

Mixture

Posterior consistency

Posterior distribution

Kullback–Leibler property

Multivariate

Density estimation

ABSTRACT

Density estimation, especially multivariate density estimation, is a fundamental problem in nonparametric inference. In the Bayesian approach, Dirichlet mixture priors are often used in practice for such problems. However, the asymptotic properties of such priors have only been studied in the univariate case. We extend the L_1 -consistency of Dirichlet mixtures in the multivariate density estimation setting. We obtain such a result by showing that the Kullback–Leibler property of the prior holds and that the size of the sieve in the parameter space in terms of L_1 -metric entropy is not larger than the order of n . However, it seems that the usual technique of choosing a sieve by controlling prior probabilities is unable to lead to a useful bound on the metric entropy required for the application of a general posterior consistency theorem for the multivariate case. We overcome this difficulty by using a structural property of Dirichlet mixtures. Our results apply to a multivariate normal kernel even when the multivariate normal kernel has a general variance–covariance matrix.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Density estimation, especially multivariate density estimation, is a fundamental problem in nonparametric inference. It serves as the basis of many other statistical methods, including semi-parametric regression, nonparametric regression [9], clustering, discriminant analysis [3] and robust estimation [15]. A Bayesian approach to density estimation often uses Dirichlet mixtures as a prior, as in [8]. On the space \mathbb{R}^d , the d -dimensional normal density function

$$\phi_d(x, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} x^T \Sigma^{-1} x \right\}$$

with mean $(0, \dots, 0)$ and variance–covariance matrix Σ is often chosen to be the kernel function. West et al. [14] developed an algorithm to calculate posterior distributions for Dirichlet mixture priors with multivariate normal density as the kernel. There are some other algorithms in the literature, such as that of Ormoneit and Tresp [10], for Bayesian estimation of Gaussian mixtures.

We formulate Dirichlet mixture models in detail and introduce the notations used in this paper as follows. Let \mathcal{F} be the space of all densities on \mathbb{R}^d with respect to Lebesgue measure. Let Π be a prior on \mathcal{F} . Given $f \in \mathcal{F}$, let X_1, X_2, \dots, X_n denote

[☆] This research was partially supported by NSF grant DMS 0349111.

* Corresponding author.

E-mail addresses: yw392@cornell.edu, feng.y.wu@gmail.com (Y. Wu), sghosal@stat.ncsu.edu (S. Ghosal).

a set of d -dimensional observations, which are independent and identically distributed (i.i.d.) with the density function f . Also, let \mathbb{X}_n stand for (X_1, X_2, \dots, X_n) . Let F and F_0 stand for the probability measure corresponding respectively to density function f and f_0 . We denote $[-a, a] \times \dots \times [-a, a]$ by B_a . Though any parametric family of probability densities can be considered as a kernel in Dirichlet mixtures, we restrict our attention to the multivariate normal kernel in this paper. Let $\phi_d(x, \Sigma)$ be as described before, and let $\phi(x, \sigma^2)$ denote $\phi_1(x, \sigma^2)$, the one-dimensional normal density function with mean 0 and standard deviation σ . Let $\lambda_1(\Sigma), \dots, \lambda_d(\Sigma)$ be the eigenvalues of the matrix Σ in the order from the smallest to the largest. Let $\Theta = \mathbb{R}^d$ and \mathcal{M} be the set of probability measures on Θ . If $P \in \mathcal{M}$, then $f_{\Sigma, P}$ will stand for the convolution of ϕ_Σ and P , i.e.,

$$f_{\Sigma, P}(x) = (\phi_d(\cdot, \Sigma) * P)(x) = \int \phi_d(x - \theta, \Sigma) dP(\theta). \quad (1)$$

We consider a prior μ for Σ and a prior Π^* on \mathcal{M} . The prior $\mu \times \Pi^*$ through the map $(\Sigma, P) \mapsto f_{\Sigma, P}$ induces a prior on \mathcal{F} . We denote this prior by Π . Thus $(\Sigma, P) \sim \mu \times \Pi^*$ and given (Σ, P) , X_1, X_2, \dots, X_n are i.i.d. $f_{\Sigma, P}$.

Recall that an L_1 -neighborhood of probability density f_0 is a set containing $\{f \in \mathcal{F} : \|f - f_0\| < \epsilon\}$, where $\|f - f_0\| = \int |f(x) - f_0(x)| dx$. By L_1 -consistency of the posterior at f_0 , we mean that $\Pi(V|\mathbb{X}_n) \rightarrow 1$ either in P_{f_0} -probability or almost surely (a.s.) $[P_{f_0}]$ for any L_1 -neighborhood V of f_0 . Note that the Hellinger distance is equivalent to the L_1 -distance, so “Hellinger consistency” is equivalent to L_1 -consistency of the posterior. For any $f_0 \in \mathcal{F}$, we denote by $K_\epsilon(f_0)$, the Kullback–Leibler neighborhood $\{f : \int f_0 \log(f_0/f) < \epsilon\}$. We say that the Kullback–Leibler (KL) property holds at f_0 , or f_0 is in the KL support of Π , if $\Pi(K_\epsilon(f_0)) > 0$ for all $\epsilon > 0$.

For the Dirichlet mixture of multivariate normal priors as described above, however, asymptotic properties have not been studied. For the univariate case, L_1 -posterior consistency and rate of convergence results for Dirichlet mixtures of a univariate normal kernel have been thoroughly studied in the literature. For the univariate normal mixture, Ghosal et al. [4] gave conditions under which the consistency of the Dirichlet mixture models for estimating univariate density functions will hold. Tokdar [12] significantly weakened their conditions for consistency, especially if the true density is not compactly supported. Ghosal and van der Vaart [5,6] gave the rate of convergence for Bayesian univariate density estimation using Dirichlet mixtures of normal distribution as the prior.

The conditions for posterior consistency of Dirichlet mixtures in univariate density estimation are obtained by balancing the size of some sieves in the parameter space and the prior probability of the component of the sieve. Such a balancing technique cannot be applied to some widely used Dirichlet mixtures in multivariate density estimation, e.g., the Dirichlet mixture of multivariate normal densities with multivariate normal distribution as the base measure of the Dirichlet process. The technique only applies to some very restricted priors, e.g. the Dirichlet process with a base measure compactly supported or the tail mass extremely small. This is due to the fast rate of increase of metric entropies of the component of the sieve with increasing dimension; see Remark 4 in Section 4 for details. In this paper, we use a different technique to control the size of the sieve, which allows us to address L_1 -consistency in the multivariate setting with only mild restriction on the tail of the base measure of the Dirichlet process prior.

The paper is organized as follows. In Section 2, we state a theorem, which is similar to Theorem 2 of Ghosal et al. [4]. It applies to general Bayesian density estimation and is the key result towards the L_1 -consistency for Dirichlet mixtures in the multivariate setting. In Section 3, we give sufficient conditions under which the Kullback–Leibler property holds. By a theorem of Schwartz [11], this implies weak consistency. Note that Theorem 5 of Wu and Ghosal [16] for the Dirichlet mixture with a scaled type multivariate normal density as its kernel is a special case of Theorem 2 in this paper. In Section 4, we first state a lemma which gives bounds for metric entropies, then state another lemma which gives the sufficient conditions to satisfy Condition (A1). As a consequence of these, our main result, Theorem 3, gives sufficient conditions under which the Dirichlet mixtures are L_1 -consistent for multivariate density estimation. Finally, we show by an example that a Dirichlet mixture prior frequently used in practice is L_1 -consistent at any f_0 satisfying appropriate conditions. Some proofs will be given in the Appendix. For some examples of inconsistency, we refer to Ghosh and Ramamoorthi [7].

2. L_1 -consistency

The size of a space can be measured by the number of small balls required to cover the space. Let $\mathcal{G} \subset \mathcal{F}$. For $\delta > 0$, the L_1 -metric entropy is denoted by $\log N(\delta, \mathcal{G})$, where $N(\delta, \mathcal{G})$, the δ -covering number of \mathcal{G} in L_1 , is the minimum of all k such that there exist $f_1, f_2, \dots, f_k \in \mathcal{F}$ with the property $\mathcal{G} \subset \cup_{i=1}^k \{f : \|f - f_i\| < \delta\}$.

Theorem 1. Consider a multivariate normal mixture prior described by (1). Let f_0 belong to the KL support of Π . For any $\epsilon > 0$, if there exist constants $\eta < \epsilon/4$, $\beta < \epsilon^2/8$, $\xi_1 > 0$ and sequences $a_n, h_n \downarrow 0$, such that

$$(A1) \ E_{f_0} \Pi\{f_{\Sigma, P} : \sqrt{\lambda_1(\Sigma)} > h_n, P(B_{2a_n}^c) > \eta|\mathbb{X}_n\} \rightarrow 0,$$

$$(A2) \ \mu\{\sqrt{\lambda_1(\Sigma)} \leq h_n\} \leq e^{-n\xi_1},$$

$$(A3) \ \log N(\eta, V_n) < n\beta, \text{ where } V_n = \mathcal{F}_n \cap U^c, U = \{\|f - f_0\| < \epsilon\} \text{ and } \mathcal{F}_n = \{f_{\Sigma, P} : \sqrt{\lambda_1(\Sigma)} > h_n \text{ and } P(B_{2a_n}^c) < \eta\},$$

then $\Pi(f : \|f - f_0\| < \epsilon|\mathbb{X}_n) \rightarrow 1$ in P_{f_0} -probability.

Proof. Since $\mu\{\sqrt{\lambda_1(\Sigma)} < h_n\} \leq e^{-n\epsilon_1}$, we have $\Pi(\{f_{\Sigma,P} : \sqrt{\lambda_1(\Sigma)} < h_n\}|\mathbb{X}_n) \rightarrow 0$ a.s. $[P_{f_0}^\infty]$; see, e.g., Proposition 4.4.2 of Ghosh and Ramamoorthi [7]. Since

$$\Pi(\mathcal{F}_n^c|\mathbb{X}_n) \leq \Pi(\{f_{\Sigma,P} : \sqrt{\lambda_1(\Sigma)} \leq h_n\}|\mathbb{X}_n) + \Pi(\{f_{\Sigma,P} : \sqrt{\lambda_1(\Sigma)} > h_n, P(B_{2a_n}^c) > \eta\}|\mathbb{X}_n),$$

by Condition (A1), we have $\Pi(\mathcal{F}_n^c|\mathbb{X}_n) \rightarrow 0$ in $P_{f_0}^\infty$. By Lemma 3 of the Appendix, we have $\Pi(\|f - f_0\| < \epsilon|\mathbb{X}_n) \rightarrow 1$ in P_{f_0} -probability. \square

Remark 1. If Condition (A1) is strengthened to

$$\sum_{n=1}^{\infty} E_{f_0} \Pi\{f_{\Sigma,P} : \sqrt{\lambda_1(\Sigma)} > h_n, P(B_{2a_n}^c) > \eta|\mathbb{X}_n\} < \infty,$$

then the conclusion will be strengthened to $\Pi(\|f - f_0\| < \epsilon|\mathbb{X}_n) \rightarrow 1$ a.s. $[P_{f_0}^\infty]$.

3. Kullback–Leibler property

Consider the mixture model with normal density as its kernel function. In Theorem 2 below, we give conditions under which the KL property holds. Such a property plays a very important role in consistency. The result generalizes Theorem 5 of Wu and Ghosal [16].

Theorem 2. Let $f_0(x)$ be the true density. Let a prior Π^* with weak support $\mathcal{M}(\mathbb{R}^d)$ be given on the mixing distribution P , and independently a prior μ supported on the set of all non-negative definite $d \times d$ symmetric matrices is assigned to Σ . Let Π stand for the prior on \mathcal{F} induced via the map $(P, \Sigma) \mapsto f_{\Sigma,P}(x)$, where $f_{\Sigma,P}$ is a normal mixture as in (1). Assume that the following conditions hold:

- (B1) $0 < f_0(x) < M^*$ for some constant M^* and all $x \in \mathbb{R}^d$;
- (B2) $|\int f_0(x) \log f_0(x) dx| < \infty$;
- (B3) for some $\delta > 0$, $\int f_0(x) \log \frac{f_0(x)}{\phi_\delta(x)} dx < \infty$, where $\phi_\delta(x) = \inf_{\|t-x\| < \delta} f_0(t)$;
- (B4) for some $\rho > 0$, $\int \|x\|^{2(1+\rho)} f_0(x) dx < \infty$.

Then f_0 is in the KL support of the prior.

Note that the following corollary is Theorem 5 of Wu and Ghosal [16]. It applies to kernel mixture models with a multivariate normal density as the kernel function while the variance–covariance matrix for the density function is restricted to a scalar identity matrix. Therefore the prior for such a model is given on the mixing probability P and the scale parameter.

Corollary 1 (Theorem 5 of Wu and Ghosal [16]). Assume that the weak support of prior Π is $\mathcal{M}(\mathbb{R}^d \times \mathbb{R}^+)$. Let $f_0(x)$ be a continuous density on \mathbb{R}^d satisfying

- (C1) for some $0 < M < \infty$, $0 < f_0(x) \leq M$ for all $x \in \mathbb{R}^d$;
- (C2) $|\int f_0(x) \log f_0(x) dx| < \infty$;
- (C3) for some $\delta > 0$, $\int f_0(x) \log \frac{f_0(x)}{\phi_\delta(x)} dx < \infty$, where $\phi_\delta(x) = \inf_{\|t-x\| < \delta} f_0(t)$;
- (C4) for some $\rho > 0$, $\int \|x\|^{2(1+\rho)} f_0(x) dx < \infty$.

Then $f_0 \in \text{KL}(\Pi^*)$.

Proof of Theorem 2. First, we bound the multivariate normal density function with the general variance–covariance matrix Σ by a multivariate normal density function with variance–covariance matrix of the form hI_d , where I_d stands for the identity matrix of order d . For any $x \in \mathbb{R}^d$ and any real symmetric positive definite matrix Σ ,

$$\|x\|^2 / \lambda_d(\Sigma) \leq x^T \Sigma^{-1} x \leq \|x\|^2 / \lambda_1(\Sigma). \quad (2)$$

Hence,

$$\left(\frac{\lambda_1(\Sigma)}{\lambda_d(\Sigma)} \right)^{\frac{d-1}{2}} \phi_d(x, \lambda_1(\Sigma) I_d) \leq \phi_d(x, \Sigma) \leq \left(\frac{\lambda_d(\Sigma)}{\lambda_1(\Sigma)} \right)^{\frac{d-1}{2}} \phi_d(x, \lambda_d(\Sigma) I_d). \quad (3)$$

The proofs of Theorem 5 and Theorem 2 of Wu and Ghosal [16] imply that, for any $\epsilon > 0$, we have an open set $\mathcal{P} \subset \mathcal{M}$ and $H \subset \mathbb{R}^+$ such that, for any $P \in \mathcal{P}$ and $h \in H$,

$$\int f_0(x) \log \frac{f_0(x)}{\int \phi_d(x - \theta, hI_d) dP(\theta)} dx \leq \epsilon/2. \quad (4)$$

Let $(h_0, h_1) \subset H$ such that $\log(h_1/h_0) < \epsilon/(d-1)$. By (3) and (4), we have that, if $(\lambda_1(\Sigma), \lambda_d(\Sigma)) \subset (h_0, h_1)$ and $P \in \mathcal{P}$, then

$$\begin{aligned} \int f_0(x) \log \frac{f_0(x)}{\int \phi_d(x - \theta, \Sigma) dP(\theta)} dx &\leq \int f_0(x) \log \frac{f_0(x)}{\int \phi_d(x - \theta, h_0 I_d) (h_0/h_1)^{(d-1)/2} dP(\theta)} dx \\ &= \int f_0(x) \log \frac{f_0(x)}{\int \phi_d(x - \theta, h_0 I_d) dP(\theta)} dx + (d-1) \log(h_1/h_0)/2 \\ &\leq \epsilon. \end{aligned}$$

Recall that the eigenvalues of a real symmetric matrix depend continuously on the matrix; see [13] for more details. Hence, we can choose an open set \mathcal{S} containing hI_d in the space of all positive definite matrices such that, for any $\Sigma \in \mathcal{S}$, $(\lambda_1(\Sigma), \lambda_d(\Sigma)) \subset (h_0, h_1)$. \square

4. L_1 -consistency of Dirichlet mixtures

For a Dirichlet normal mixture prior, Theorem 3 below gives sufficient conditions under which the posterior is L_1 -consistent. Before this theorem, we present two lemmas. The first lemma gives the size of the parameter space measured in terms of L_1 -metric entropy. The second one calculates the posterior probability of the complement of a sieve. The latter is the key step in controlling the size of the sieve in higher dimensions.

Lemma 1. Let a and h be positive constants and $\mathcal{P}_{h,a,\eta}^M = \{f_{P,\Sigma} : P(B_a^c) < \eta, \sqrt{\lambda_1(\Sigma)} > h, \sqrt{\lambda_d(\Sigma)} < M\}$. Then

$$\log N(\eta, \mathcal{P}_{h,a,\eta}^M) \leq K^*(a/h)^d,$$

where K^* is a constant that depends on ϵ, η and M , but not a or h .

The proof of this lemma is given in the Appendix.

The following lemma, generalizing Lemma 11 of Ghosal and van der Vaart [6], gives a very important tool for bounding the posterior probability of the complement of a sieve.

Lemma 2. Let X_1, X_2, \dots, X_n be i.i.d. with true density f_0 . Assign a Dirichlet process \mathcal{D}_α prior on the space of mixing distribution P with the base measure α , which has a positive and continuous density on \mathbb{R}^d , and independently a prior μ be given to Σ on the space of the $d \times d$ symmetric positive definite matrix. Assume that μ has support in the space of Σ with eigenvalues in $[0, M]$. Suppose that there exist positive sequences $a_n \rightarrow \infty, \epsilon_n > 0, n\epsilon_n \rightarrow \infty$ and $h_n \rightarrow 0$ such that $a_n^{2r}\epsilon_n/n \rightarrow \infty, \mu\{\sqrt{\lambda_1(\Sigma)} \leq h_n\} \leq e^{-cn}$ for some constants $d < r < d(1+\rho)$ and c , plus Conditions (B3), (B4) and the following condition holds:

(B5) $h_n^{-d} n e^{-a_n^2/2M} = o(\epsilon_n \delta_n)$, where δ_n denote a lower bound for the density of $\bar{\alpha}$ on $B_{a_n+\sqrt{M}}$.

Then $E[\Pi_n\{P(B_{2a_n}^c) > \epsilon_n | \mathbb{X}_n\}] \rightarrow 0$.

Proof. Given $\theta_1, \theta_2, \dots, \theta_n$, the observation \mathbb{X}_n is independent of P . Hence,

$$\Pi(P(B_{2a_n}^c) > \epsilon_n | \mathbb{X}_n) = E(\Pi(P(B_{2a_n}^c) > \epsilon_n | \theta_1, \theta_2, \dots, \theta_n) | \mathbb{X}_n).$$

From [2],

$$P(B_{2a_n}^c) | \theta_1, \theta_2, \dots, \theta_n \sim \text{Beta}(\alpha(B_{2a_n}^c) + N(B_{2a_n}^c), \alpha(B_{2a_n}) + N(B_{2a_n})),$$

where $N(A) = \sum_{i=1}^n \mathbb{1}_{\{\theta_i \in A\}}$. By Markov's inequality,

$$\Pi(P(B_{2a_n}^c) > \epsilon_n | \mathbb{X}_n) \leq \frac{\alpha(B_{2a_n}^c) + \sum_{i=1}^n \Pr(\theta_i \in B_{2a_n}^c, \sqrt{\lambda_1(\Sigma)} > h_n | \mathbb{X}_n)}{\epsilon_n(\alpha(\mathbb{R}^d) + n)} + \Pr(\sqrt{\lambda_1(\Sigma)} \leq h_n | \mathbb{X}_n).$$

Therefore,

$$E[\Pi\{P(B_{2a_n}^c) > \epsilon_n | \mathbb{X}_n\}] \leq \frac{\alpha(B_{2a_n}^c)}{\epsilon_n(\alpha(\mathbb{R}^d) + n)} + \frac{n \cdot E[\Pr(\theta_n \in B_{2a_n}^c, \sqrt{\lambda_1(\Sigma)} > h_n | \mathbb{X}_n)]}{\epsilon_n(\alpha(\mathbb{R}^d) + n)} + E\Pr(\sqrt{\lambda_1(\Sigma)} \leq h_n | \mathbb{X}_n).$$

The first term on the right-hand side (RHS) of the above inequality converges to zero by assumption. The third term on the RHS converges to zero, as shown in the proof of Theorem 1. To complete the proof, we shall show that

$$E[\Pr(\theta_n \in B_{2a_n}^c, \sqrt{\lambda_1(\Sigma)} > h_n | \mathbb{X}_n)]/\epsilon_n \rightarrow 0 \quad \text{as } a_n \rightarrow \infty. \quad (5)$$

To this end, we let $\theta_{-n} = \theta_1, \dots, \theta_{n-1}$, $H(\theta_1, \dots, \theta_n)$ be the joint distribution of $(\theta_1, \dots, \theta_n)$, $H_n(\theta_n|\theta_{-n})$ be the conditional distribution of θ_n given θ_{-n} and $H_{-n}(\theta_{-n})$ be the marginal distribution of θ_{-n} . Bayes' formula then gives $\Pr(\theta_n \in B_{2a_n}^c, \sqrt{\lambda_1(\Sigma)} > h_n|\mathbb{X}_n) = A(\mathbb{X}_n)/B(\mathbb{X}_n)$, where $A(\mathbb{X}_n)$ is equal to

$$\int_{\sqrt{\lambda_1(\Sigma)} > h_n} \iint_{t_n \in B_{2a_n}^c} \prod_{i=1}^n \frac{e^{-\frac{1}{2}(X_i - t_i)^T \Sigma^{-1}(X_i - t_i)}}{|\Sigma|^{1/2}} dH_n(t_n|t_{-n}) dH_{-n}(t_{-n}) d\mu(\Sigma),$$

and $B(\mathbb{X}_n)$ is equal to

$$\iiint \prod_{i=1}^n |\Sigma|^{-1/2} e^{-\frac{1}{2}(X_i - t_i)^T \Sigma^{-1}(X_i - t_i)} dH_n(t_n|t_{-n}) dH_{-n}(t_{-n}) d\mu(\Sigma).$$

We upper bound $E_{f_0} \left[\frac{A(\mathbb{X}_n)}{B(\mathbb{X}_n)} \right] / \epsilon_n$ by splitting it into two parts:

$$\epsilon_n^{-1} E_{f_0} \left[\frac{A(\mathbb{X}_n)}{B(\mathbb{X}_n)} \right] \leq \epsilon_n^{-1} \sup_{\mathbb{X}_n \in B_{a_n}^n} \left(\frac{A(\mathbb{X}_n)}{B(\mathbb{X}_n)} \right) \cdot \Pr(\mathbb{X}_n \in B_{a_n}^n) + \epsilon_n^{-1} \int_{\mathbb{X}_n \notin B_{a_n}^n} \frac{A(\mathbb{X}_n)}{B(\mathbb{X}_n)} f_0^n(\mathbb{X}_n) d\mathbb{X}_n. \quad (6)$$

We compute the first term on the RHS of (6) first. To this end, we lower bound $B(\mathbb{X}_n)$ when $\mathbb{X}_n \in B_{a_n}^n$. Observe that the conditional distribution $H_n(\cdot|\theta_{-n})$ of θ_{-n} can be structurally described as

$$\theta_n|\theta_{-n} = \begin{cases} \theta_i, & \text{with probability } 1/(\alpha(\mathbb{R}^d) + n - 1), \quad i = 1, \dots, n - 1, \\ \sim \tilde{\alpha}, & \text{with probability } \alpha(\mathbb{R}^d)/(\alpha(\mathbb{R}^d) + n - 1). \end{cases}$$

Therefore, with δ_n as defined in Lemma 2, we have

$$\begin{aligned} \int e^{-\frac{1}{2}(X_n - t_n)^T \Sigma^{-1}(X_n - t_n)} dH_n(t_n|t_{-n}) &\geq \frac{\alpha(\mathbb{R}^d)}{\alpha(\mathbb{R}^d) + n - 1} \int e^{-\frac{1}{2}(X_n - t_n)^T \Sigma^{-1}(X_n - t_n)} d\tilde{\alpha}(t_n) \\ &\geq \frac{\alpha(\mathbb{R}^d)}{\alpha(\mathbb{R}^d) + n - 1} \int_{(X_n - t_n)^T \Sigma^{-1}(X_n - t_n) \leq 1} e^{-\frac{1}{2}(X_n - t_n)^T \Sigma^{-1}(X_n - t_n)} d\tilde{\alpha}(t_n) \\ &\geq \frac{\alpha(\mathbb{R}^d)}{\alpha(\mathbb{R}^d) + n - 1} e^{-1/2} \delta_n 2^d |\Sigma|^{1/2}. \end{aligned}$$

This leads to

$$B(\mathbb{X}_n) \geq \frac{\alpha(\mathbb{R}^d) e^{-\frac{1}{2}} \delta_n 2^d}{\alpha(\mathbb{R}^d) + n - 1} \iint \frac{\prod_{i=1}^{n-1} \exp \left[-\frac{1}{2}(X_i - t_i)^T \Sigma^{-1}(X_i - t_i) \right]}{|\Sigma|^{\frac{n-1}{2}}} dH_{-n}(t_{-n}) d\mu(\Sigma), \quad (7)$$

for all $\mathbb{X}_n \in B_{a_n}^n$.

Now, for $\mathbb{X}_n \in B_{a_n}^n$, an upper bound for $A(\mathbb{X}_n)$ will be obtained. Note that, for $X \in B_{a_n}$ and $t \in B_{2a_n}^c$, when n is large such that $a_n > M^{1/2}$ and $\sqrt{\lambda_1(\Sigma)} > h_n$, we have

$$|\Sigma|^{-1/2} \exp \left[-\frac{1}{2}(X - t)^T \Sigma^{-1}(X - t) \right] \lesssim h_n^{-d} e^{-\frac{a_n^2}{2M}}.$$

Therefore,

$$A(\mathbb{X}_n) \leq h_n^{-d} e^{-\frac{a_n^2}{2M}} \iint \frac{\prod_{i=1}^{n-1} \exp \left[-\frac{1}{2}(X_i - t_i)^T \Sigma^{-1}(X_i - t_i) \right]}{|\Sigma|^{\frac{n-1}{2}}} dH_{-n}(t_{-n}) d\mu(\Sigma), \quad (8)$$

for all $\mathbb{X}_n \in (B_{a_n})^n$. Combining (7) and (8), we have

$$\epsilon_n^{-1} \sup_{\mathbb{X}_n \in (B_{a_n})^n} \left(\frac{A(\mathbb{X}_n)}{B(\mathbb{X}_n)} \right) \cdot \Pr(\mathbb{X}_n \in B_{a_n}^n) \leq \left(\frac{h_n^{-d} e^{-\frac{a_n^2}{2M}}}{e^{-\frac{1}{2}} \delta_n 2^d} \right) / \left(\frac{\alpha(\mathbb{R}^d)}{\alpha(\mathbb{R}^d) + n - 1} \epsilon_n \right) \rightarrow 0, \quad (9)$$

for all $\mathbb{X}_n \in B_{a_n}^n$ by Condition (B5).

Now, we compute the second term in (6). Obviously, we have $\frac{A(\mathbb{X}_n)}{B(\mathbb{X}_n)} \leq 1$, and

$$\int_{\mathbb{X}_n \notin B_{a_n}^n} \frac{A(\mathbb{X}_n)}{B(\mathbb{X}_n)} f_0^n(\mathbb{X}_n) d\mathbb{X}_n \leq \int_{X \in B_{a_n}^c} n f_0(X) dX. \quad (10)$$

Hence, if

$$P_{f_0}(X \in B_{a_n}^c) = o(n^{-1}\epsilon_n), \quad (11)$$

then by (9) and (10), (5) holds. By Condition (B4) and the Markov inequality,

$$\begin{aligned} P_{f_0}(X \in B_{a_n}^c) &\leq P_{f_0}(\|X\| > a_n) = P_{f_0}(\|X\|^{2(1+\rho)d} > a_n^{2(1+\rho)d}) \\ &\leq \frac{E(\|X\|^{2(1+\rho)d})}{a_n^{2(1+\rho)d}}. \end{aligned}$$

Therefore, choosing r such that $d < r < (1 + \rho)d$, we have $P_{f_0}(X \in B_{a_n}^c) = o(a_n^{-2r})$. Now, by the assumption that $a_n^{2r}\epsilon_n/n \rightarrow \infty$, we have that (11) holds. \square

Remark 2. To replace the convergence in probability by a.s. convergence in the result of this lemma, we need to replace Condition (B5) by the following.

(B5') $\sum_{i=1}^{\infty} h_i^{-d} i e^{-a_i^2/(2M)} / (\epsilon_i \delta_i) < \infty$, and assume that $a_n^{2r}\epsilon_n/n^{1+\rho} \rightarrow \infty$ for some $\rho > 0$.

Now we have our main theorem by combining the above results with ϵ_n taken to be fixed in Lemma 2.

Theorem 3. For a prior Π described as in Lemma 2, if, for any $\epsilon > 0$ and any $\beta < \epsilon^2/8$, there exist sequences $a_n \rightarrow \infty$, $a_n^{2r}/n \rightarrow \infty$, $h_n \downarrow 0$ and $\beta_1 > 0$, such that Conditions (B1)–(B5) and the following conditions hold:

(B6) $\mu\{\sqrt{\lambda_1(\Sigma)} \leq h_n\} \leq e^{-n\beta_1}$;

(B7) $(a_n/h_n)^d < n\beta$,

then $\Pi(f : \|f - f_0\| < \epsilon | \mathbb{X}_n) \rightarrow 1$ in P_{f_0} -probability.

If (B5) is strengthened to (B5'), then $\Pi(f : \|f - f_0\| < \epsilon | \mathbb{X}_n) \rightarrow 1$ a.s. $[P_{f_0}^\infty]$.

Remark 3. In the above theorem, we consider the kernel function to be a multivariate normal density function with a general variance–covariance matrix Σ , and a prior μ is given for Σ . Note that the model with kernel function being the multivariate normal density function with variance–covariance matrix hI_d , where I_d is the d -dimensional identity matrix and h is given a prior μ , is a special case covered by this theorem.

Remark 4. One of the major differences between Theorem 3 in this paper and Theorem 7 of Ghosal et al. [4] is that we only require $E[\Pi(\{P(B_{2a_n}^c) > \epsilon_n\} | \mathbb{X}_n)] \rightarrow 0$ here, while a stronger condition for the prior on the space of the mixing distribution P ,

$$\mathcal{D}_\alpha\{P[-a_n, a_n] < 1 - \delta\} < e^{-n\beta_0} \quad (12)$$

for some $\beta_0 > 0$, was used for the univariate case. However, such a condition cannot be satisfied for many common choice of α in the multivariate cases. For example, if α is chosen as multivariate normal, then a_n must be at least the order of \sqrt{n} to satisfy (12). With $h_n \downarrow 0$, now Condition (B7) cannot be satisfied for this choice of a_n .

Now we give an example of a prior for which L_1 -consistency holds. We show this in the following corollary by applying the theorem above.

First, we define a distribution that will be used as the prior for the variance–covariance matrix of the multivariate normal density kernel. Let $W(H, q)$ denote the Wishart distribution with scale matrix H and degree of freedom $q \geq d$, an integer. Let \mathcal{S} denote the set of all variance–covariance matrices and $\mathcal{S}_M \subset \mathcal{S}$ denote the subset of all $d \times d$ positive definite matrices with $\text{tr}(\Sigma) < M$. Hence, for all $\Sigma \in \mathcal{S}_M$, $\lambda_d(\Sigma) < M$. Let μ^* denote the $W(H, q)$ distribution restricted to \mathcal{S}_M ; that is,

$$\Pr_{\mu^*}(\Sigma \in \mathcal{T}) = \frac{P_{W(H, q)}(\{\Sigma \in \mathcal{T}\} \cap \{\Sigma \in \mathcal{S}_M\})}{P_{W(H, q)}(\{\Sigma \in \mathcal{S}_M\})}.$$

If $\Sigma \sim \mu^*$, we shall say that Σ follows a truncated Wishart distribution with parameters (H, q, M) .

Corollary 2. Assume that f_0 satisfies Conditions (B1)–(B4). Let the prior be as described in Theorem 3, where α is chosen such that $\bar{\alpha} := \alpha/\alpha(\mathbb{R}^d)$ is a normal distribution with variance $\sigma^2 I_d$, $\sigma^2 > dM$ and $\Sigma^{-k} \sim \mu^*$, a truncated Wishart distribution with parameter $(\lambda I_d, q, M)$, where $d < k < r$ and $r \in (d, (1 + \rho)d)$ as specified in Lemma 2. Then $\Pi(\|f - f_0\| < \epsilon | \mathbb{X}_n) \rightarrow 1$ a.s. $[P_{f_0}^\infty]$.

Proof. We need to show that there exist sequences $\{a_n\}$ and $\{h_n\}$ such that Conditions (B5')–(B7) are satisfied. Let $a_n = C_1 n^{1/(2k)}$ and $h_n = C_2 n^{-1/(2k)}$, such that $(C_1/C_2)^d < \beta < \epsilon^2/8$. Condition (B7) is obviously satisfied. Condition (B5') is satisfied by choosing $\sigma^2 > dM$, so $d/(2\sigma^2) - 1/(2M) < 0$ and hence $\sum_{i=1}^{\infty} h_i^{-d} i \exp[(d/(2\sigma^2) - 1/(2M))a_i^2] < \infty$. To see that Condition (B6) is satisfied, we have

$$\begin{aligned} \mu^* \{\sqrt{\lambda_1(\Sigma)} \leq h_n\} &= \mu^* \{\lambda_d(\Sigma^{-k}) \geq h_n^{-2k}\} \\ &\leq \mu^* \{\text{tr}(\Sigma^{-k}) \geq h_n^{-2k}\}. \end{aligned}$$

By the definition of a Wishart distribution, if $T \sim W(\lambda I_d, q)$, then $\text{tr}(T) \sim \lambda \chi_{dq}^2$. Since a χ^2 distribution has exponential tail, with $V \sim \chi_{dq}^2$,

$$\mu^* \{\text{tr}(\Sigma^{-k}) \geq h_n^{-2k}\} \leq \Pr(V \geq \lambda^{-1} C_2^{-2k} n) \frac{1}{P_{W(I_d, q)}(\{\Sigma \in \mathcal{S}_M\})} \lesssim e^{-cn},$$

for some constant $c > 0$. Finally, observe that $a_n^{2r}/n = n^{r/k-1} \rightarrow \infty$ if $r > k$. \square

Appendix

Lemma 3. Let Π be a prior on \mathcal{F} . Suppose that $f_0 \in \mathcal{F}$ is in the KL support of Π . Let $U = \{f : \|f - f_0\| < \epsilon\}$. If there is $\eta < \epsilon/4$, $\beta < \epsilon^2/8$ and $\mathcal{F}_n \subset \mathcal{F}$ such that, for all n sufficiently large,

- (i) $\Pi(\mathcal{F}_n^c | \mathbb{X}_n) \xrightarrow{p_{f_0}} 0$,
- (ii) $\log N(\eta, V_n) < n\beta$,

then $\Pi(U^c | X_1, \dots, X_n) \xrightarrow{p_{f_0}} 0$.

To prove this lemma, we use a result of Barron [1], which is slightly differently stated in the following lemma.

Lemma 4. Let Π be a prior on \mathcal{F} , $f_0 \in \mathcal{F}$ be in the KL support of Π and U_n be a sequence of neighborhoods of f_0 . Then (i) and (ii) below are equivalent.

- (i) $\Pi(U_n^c | \mathbb{X}_n) \xrightarrow{p_{f_0}^\infty} 0$.
- (ii) There exist subsets V_n, W_n of \mathcal{F} , and a sequence of tests $\{\varphi_n(\mathbb{X}_n)\}$ such that
 - (a) $U_n^c \subset V_n \cup W_n$,
 - (b) $\Pi(W_n | \mathbb{X}_n) \xrightarrow{p_{f_0}^\infty} 0$,
 - (c) there exists a sequence of tests, such that $\varphi_n(\mathbb{X}_n) \xrightarrow{p_{f_0}^\infty} 0$ and $\inf_{f \in V_n} E_f \varphi_n \geq 1 - ce^{-n\beta}$, where c and β are positive.

The proof of Lemma 3 follows along almost the same lines as those in the proof of Theorem 2 in Ghosal et al. [4]. The only difference is that we verify the conditions of Lemma 4 here instead of the result of Barron [1] in its original form.

Proof of Lemma 1. We prove this lemma through the following three lemmas.

Lemma 5. For any $\epsilon, a > 0$ and Σ positive definite,

$$\log N(2\epsilon, \mathcal{F}_{\Sigma, a}) \leq \left(\sqrt{\frac{8d}{\pi}} \frac{a}{\sqrt{\lambda_1(\Sigma)}\epsilon} + 1 \right)^d \left\{ 1 + \log \left(\frac{1+\epsilon}{\epsilon} \right) \right\},$$

where $\mathcal{F}_{\Sigma, a} = \{f_{p, \Sigma} : P(B_a) = 1\}$.

Proof. Let A be an orthogonal matrix such that

$$A \Sigma^{-1} A^T = \text{diag} \left(\frac{1}{\lambda_1(\Sigma)}, \dots, \frac{1}{\lambda_d(\Sigma)} \right)$$

and

$$L = \text{diag}(\sqrt{\lambda_1(\Sigma)}, \dots, \sqrt{\lambda_d(\Sigma)}).$$

For $\theta_1 \neq \theta_2$,

$$\begin{aligned}
 \|\phi_{\theta_1, \Sigma} - \phi_{\theta_2, \Sigma}\| &= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \left| e^{-\frac{1}{2}(x-\theta_1)^T \Sigma^{-1}(x-\theta_1)} - e^{-\frac{1}{2}(x-\theta_2)^T \Sigma^{-1}(x-\theta_2)} \right| dx \\
 &= \int_{\mathbb{R}^d} \frac{\left| e^{-\frac{[A(x-\theta_1)]^T A \Sigma^{-1} A^T A(x-\theta_1)}{2}} - e^{-\frac{[A(x-\theta_2)]^T A \Sigma^{-1} A^T A(x-\theta_2)}{2}} \right|}{(2\pi)^{d/2} |\Sigma|^{1/2}} d(Ax) \\
 &= \int_{\mathbb{R}^d} \frac{\left| e^{-\frac{[L^{-1}A(x-\theta_1)]^T L^{-1}A(x-\theta_1)}{2}} - e^{-\frac{[L^{-1}A(x-\theta_2)]^T L^{-1}A(x-\theta_2)}{2}} \right|}{(2\pi)^{d/2}} d(L^{-1}Ax) \\
 &= \int_{\mathbb{R}^d} \frac{\left| e^{-\frac{1}{2}(y-L^{-1}A\theta_1)^T (y-L^{-1}A\theta_1)} - e^{-\frac{1}{2}(y-L^{-1}A\theta_2)^T (y-L^{-1}A\theta_2)} \right|}{(2\pi)^{d/2}} dy \\
 &= (2\pi)^{-d/2} \int \left| e^{-\frac{1}{2}\|z-\alpha\|^2} - e^{-\frac{1}{2}\|z\|^2} \right| dz \\
 &= (2\pi)^{-d/2} \left\{ \int_{\alpha^T z > \frac{\|\alpha\|^2}{2}} e^{-\frac{1}{2}\|z-\alpha\|^2} dz - \int_{\alpha^T z > \frac{\|\alpha\|^2}{2}} e^{-\frac{1}{2}\|z\|^2} dz \right. \\
 &\quad \left. + \int_{\alpha^T z \leq \frac{\|\alpha\|^2}{2}} e^{-\frac{1}{2}\|z\|^2} dz - \int_{\alpha^T z \leq \frac{\|\alpha\|^2}{2}} e^{-\frac{1}{2}\|z-\alpha\|^2} dz \right\} \\
 &= (2\pi)^{-d/2} \left\{ \int_{\alpha^T u > -\frac{\|\alpha\|^2}{2}} e^{-\frac{1}{2}\|u\|^2} du - \int_{\alpha^T u > -\frac{\|\alpha\|^2}{2}} e^{-\frac{1}{2}\|u\|^2} du \right. \\
 &\quad \left. + \int_{\alpha^T u \leq \frac{\|\alpha\|^2}{2}} e^{-\frac{1}{2}\|u\|^2} du - \int_{\alpha^T u \leq \frac{\|\alpha\|^2}{2}} e^{-\frac{1}{2}\|u\|^2} du \right\} \\
 &= 4 \Pr \left\{ 0 \leq \alpha^T Z \leq \frac{\|\alpha\|^2}{2} \right\} \\
 &= 4 \left\{ \Phi \left(\frac{\|\alpha\|}{2} \right) - \Phi(0) \right\} \\
 &\leq \sqrt{\frac{2}{\pi}} \|\alpha\| \leq \sqrt{\frac{2}{\pi}} \frac{\|\theta_1 - \theta_2\|}{\sqrt{\lambda_1(\Sigma)}},
 \end{aligned}$$

where $\alpha = L^{-1}A(\theta_1 - \theta_2)$ and $Z \sim \text{Norm}(0, I_d)$. Given ϵ , let k be the smallest integer greater than $\frac{\sqrt{8a}\sqrt{d}}{\sqrt{\pi}\sqrt{\lambda_1(\Sigma)}\epsilon}$. Divide $(-a, a]^d$ into $N = k^d$ cubes. Let $E_{i_1, i_2, \dots, i_d} = \prod_{j=1}^d (-a + \frac{2a(i_j-1)}{k}, -a + \frac{2ai_j}{k}]$, where $1 \leq i_1, i_2, \dots, i_d \leq k$. If θ and θ' belong to the same cube, $\|\phi_d(\theta, \Sigma) - \phi_d(\theta', \Sigma)\| < \epsilon$. Let $\mathcal{P}_N = \{(P_1, \dots, P_N) : P_i > 0, \sum_{i=1}^N P_i = 1\}$ be the N -dimensional probability simplex, and let \mathcal{P}_N^* be a ϵ -net in \mathcal{P}_N ; that is, given $P \in \mathcal{P}_N$, there is $P^* = (P_1^*, \dots, P_N^*) \in \mathcal{P}_N^*$ such that $\sum_{i=1}^N |P_i - P_i^*| < \epsilon$. Let $\mathcal{F}^* = \{\sum_{i=1}^N P_i^* \phi_{\theta_i, \Sigma} : (P_1^*, \dots, P_N^*) \in \mathcal{P}_N^*\}$. We shall show that \mathcal{F}^* is a 2ϵ -net in $\mathcal{F}_{h,a}$. If $f_{P, \Sigma} = \phi_{\Sigma} * P \in \mathcal{F}_{h,a}$, set $P_i = P(E_i)$. Let $(P_1^*, \dots, P_N^*) \in \mathcal{P}_N^*$ such that $\sum_{i=1}^N |P_i - P_i^*| < \epsilon$. Then

$$\begin{aligned}
 \left\| \int \phi_{\theta, \Sigma} dP(\theta) - \sum_{i=1}^N P_i^* \phi_{\theta_i, \Sigma} \right\| &\leq \left\| \int \phi_{\theta, \Sigma} dP(\theta) - \sum_{i=1}^N \int \mathbb{1}_{E_i}(\theta) \phi_{\theta_i, \Sigma} dP(\theta) \right\| + \left\| \sum_{i=1}^N P_i \phi_{\theta_i, \Sigma} - \sum_{i=1}^N P_i^* \phi_{\theta_i, \Sigma} \right\| \\
 &\leq 2\epsilon.
 \end{aligned}$$

This shows that $N(2\epsilon, \mathcal{F}_{h,a}) \leq N(\epsilon, \mathcal{P}_N)$. The covering number of \mathcal{P}_N is bounded by $(N/\epsilon)^N (1 + \epsilon)^N \frac{1}{N!}$; see Lemma 1 of Ghosal et al. [4]. So,

$$\log N(2\epsilon, \mathcal{F}_{\Sigma,a}) \leq N \left(1 + \log \frac{1 + \epsilon}{\epsilon} \right) \leq \left(\frac{\sqrt{8a}\sqrt{d}}{\sqrt{\pi}\sqrt{\lambda_1(\Sigma)}\epsilon} + 1 \right)^d \left(1 + \log \frac{1 + \epsilon}{\epsilon} \right). \quad \square$$

The following two lemmas are similar to Lemmas 2 and 3 of Ghosal et al. [4], respectively, and can be proved along similar lines.

Lemma 6. Let $\mathcal{F}_{\Sigma,a,\epsilon} = \{f_{\Sigma,p} : P(B_a) > 1 - \epsilon\}$. Then $N(3\epsilon, \mathcal{F}_{\Sigma,a,\epsilon}) \leq N(\epsilon, \mathcal{F}_{\Sigma,a})$.

Lemma 7. Let $M > 0$ and let $\mathcal{F}_{h,a,\epsilon}^M = \cup\{\mathcal{F}_{\Sigma,a,\epsilon} : \lambda_1(\Sigma) > h, \lambda_d(\Sigma) < M\}$. If $a > \sqrt{d}M/\sqrt{\epsilon}$, then $\mathcal{F}_{h,a,\epsilon}^M \subset \mathcal{F}_{hd,2a,2\epsilon}$. \square

References

- [1] A.R. Barron, The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions, Technical Report 7, Dept. Statistics, Univ. Illinois, Champaign, 1988.
- [2] T.S. Ferguson, A Bayesian analysis of some nonparametric problems, *Ann. Statist.* 1 (1973) 209–230.
- [3] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, *J. Amer. Statist. Assoc.* 97 (2002) 611–631.
- [4] S. Ghosal, J.K. Ghosh, R.V. Ramamoorthi, Posterior consistency of Dirichlet mixtures in density estimation, *Ann. Statist.* 27 (1999) 143–158.
- [5] S. Ghosal, A.W. van der Vaart, Entropies and rates of convergence for Bayes and maximum likelihood estimation for mixture of normal densities, *Ann. Statist.* 29 (2001) 1233–1263.
- [6] S. Ghosal, A.W. van der Vaart, Posterior convergence rates of Dirichlet mixtures at smooth densities, *Ann. Statist.* 35 (2007) 697–723.
- [7] J.K. Ghosh, R.V. Ramamoorthi, *Bayesian Nonparametrics*, Springer-Verlag, New York, 2003.
- [8] A.Y. Lo, On a class of Bayesian nonparametric estimates I: density estimates, *Ann. Statist.* 12 (1984) 351–357.
- [9] P. Müller, A. Erkanli, M. West, Bayesian curve fitting using multivariate normal mixtures, *Biometrika* 83 (1996) 67–79.
- [10] D. Ormoneit, V. Tresp, Improved Gaussian mixture density estimates using Bayesian penalty terms and network averaging, in: *Advances in Neural Information Processing*, vol. 8, MIT Press, 1996, pp. 542–548.
- [11] L. Schwartz, On Bayes procedures, *Z. Wahrsch. Verw. Gebiete* 4 (1965) 10–26.
- [12] S. Tokdar, Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression, *Sankhya* 67 (2006) 90–110.
- [13] E.E. Tyrtshnikov, *A Brief Introduction to Numerical Analysis*, Birhäuser, Boston, 1997.
- [14] M. West, P. Müller, M.D. Escobar, Hierarchical priors and mixture models, with applications in regression and density estimation, in: P.R. Freeman, A.F.M. Smith (Eds.), *Aspects of Uncertainty: A Tribute to D.V. Lindley*, Wiley, Chichester, 1994, pp. 363–386.
- [15] Mi-Ja Woo, T.N. Sriram, Robust estimation of mixture complexity for count data, *Computational Statistics & Data Analysis* 51 (2007) 4379–4392.
- [16] Y. Wu, S. Ghosal, Kullback–Leibler property of kernel mixture priors in Bayesian density estimation, *Electronic J. Statist.* 2 (2008) 298–331; Correction: *Electronic J. Statist.* 3 (2009) 316–317.